Visual Question Answering and Image Caption Generation Using Deep Learning

B Lalitha Rajeswari[1], M Keerthi[2], V V A L Sravya[3], K Pradeep Choudhary[4], K Govinda Reddy[5]
[1]Supervisor, Assistant Professor, Computer Science and Engineering Artificial Intelligence and Machine learning, Vasireddy Venkatadri Institute of Technology, Andhra Pradesh, India
[2,3,4,5]Student, Computer Science and Engineering - Artificial Intelligence and Machine learning, Vasireddy Venkatadri Institute of Technology, Andhra Pradesh, India

***Abstract*—The study develops two computer vision applications through AI technology which include Visual Question Answering (VQA) and Image Caption Generation. A VQA system operates by combining computational vision methods with natural lan- guage processing capabilities to both read pictures and produce appropriate user responses. The system makes use of ResNet-50 to extract image features and BERT to process textual questions thus providing precise context-based answers. The data augmentation process benefits from the similar image search API and users gain a better experience through JavaScript-based image enhancement.**

With deep learning methods the Image Caption Generation system produces meaningful text descriptions that describe im- ages. The system integrates VGG16 for visual feature extraction method together with LSTM-based sequence modeling to produce natural language captions. The system receives education through extensive databases containing image-caption pairs that allow it to create coherent captions that align with the visual context. The system usability receives enhancement through additional features which include similar image retrieval and image en- hancement techniques.

The training methods include Adam optimization and Cosine Annealing Learning Rate Scheduler and Early Stopping to achieve both accuracy and efficient learning in both models. Real-world performance assessments indicate that VQA achieved 90.11

***Index Terms*—Visual Question Answering, Image Caption Generation, ResNet-50, BERT, VGG16, LSTM,**

I. INTRODUCTION

Deep learning technology transformed different business sectors through rapid artificial intelligence advancements pri- marily in computer vision and natural language processing (NLP) domains. The efficient analysis of both textual and visual content by machines results from these technological advancements leading to improved AI solutions. Visual Ques- tion Answering (VQA) and Image Caption Generation have transformed into essential applications because they let AI systems answer and generate text descriptions from images through vision and language integration.

The AI process of Visual Question Answering (VQA) requires machines to analyze images while creating answers that stem from image content. Both image processing to obtain visual details and natural language understanding methods to analyze questions combine to accomplish this task. The VQA system combines ResNet-50 as a pretrained deep learning model to extract thorough image features together with BERT (Bidirectional Encoder Representations from Transformers) for text encoding operations. The collected features unite for classification purposes to create the proper response. User experience and overall performance receive significant benefits from the similar image search and JavaScript-based image enhancement because both functionalities enhance dataset augmentation and image clarity.

aci'on de la Generacio'n de Captiones de Ima'genes *functions* as a deep learning-based application which generates textual descriptions automatically from images. The system links CNNs for visual features extraction to RNNs with specialized LSTM networks for producing sequential text results. Large datasets made up of images and their accompanying cap- tions serve for training the model which subsequently learns contextual understanding and produces captions that contain proper structure. The system performance improves through the integration of image similarity retrieval elements in con- junction with JavaScript-based image enhancement techniques that enhance feature extraction and lead to better caption accuracy.

The VQA model together with Image Caption Generation employs Adam optimization and Cosine Annealing Learning Rate Scheduler and Early Stopping approaches to deliver efficient learning while achieving optimal performance. These techniques enable the models to adjust successfully and min- imize overfitting which leads to better accuracy results. Both VQA model along with Image Caption Generation model demonstrated practical success through their respective accu- racy rates of 90.11

AI-driven models find practical uses in four different in- dustrial sectors - automated customer support and assistive technologies for visual impairment combined with intelligent content search tools accompanied by automated metadata processing capabilities. The combination of deep learning advances and computational efficiency improvements will re- sult in better performance for real-world application of these models. The combination of vision and language-based AI models in this research helps create better human-machine interactions through systems that process and analyze human- like responses.

II. LITERATURE SURVEY

Deep learning has revolutionized computer vision and nat- ural language processing (NLP) through new applications like Visual Question Answering (VQA) and Image Caption Gener- ation during the past few years. These technological processes enable machines to analyze images after processing them for meaningful response generation or descriptive outputs which serves as a foundation for artificial intelligence development.

Visual Question Answering (VQA) grew from it begin- nings using convolutional neural networks (CNNs) for image task feature extraction and long short-term memory (LSTM) networks to process textual queries according to antol2015vqa. The implementation of attention mechanisms transformed the performance of these models because they enabled the system to pay attention to important visual elements when answering questions [2]. BERT represents one of the transformer-based architectures that improved understanding of complex textual queries and provided context-aware response capabilities [3]. The combination of image and text embeddings through biomodal fusion has gained attention with deep learning methods like bilinear pooling and self-attention mechanisms producing better response accuracy [4].

The process of Image Caption Generation has experienced parallel advancements. Researchers at first used template- based captioning methods to produce stiff and artificial de- scriptions [5]. The development of deep learning introduced CNN-RNN architectures that combined VGG16 or Incep- tionV3 image feature extraction with LSTM networks which produced textual descriptions according to xu2015show. A new generation of attention-based captioning mechanisms al- lowed models to develop dynamic focus on specific image parts when producing captions [7]. The use of transformers for captioning has brought about better fluency and improved grammatical accuracy based on recent studies [8].

Various extensive datasets including MS COCO and both Flickr30k and Visual Genome have proved essential for train- ing and testing VQA and image captioning models [9]. Voice queries benefit from three popular pretrained word embedding systems known as Word2Vec, GloVe and FastText which improve textual representations significantly in these tasks [10]. Researchers integrated reinforcement learning techniques to both improve the correctness and increase the diversity of generated captions [11].

The combination between ResNet-50 image feature extrac- tion and transformer analysis achieves cutting-edge results for both VQA tasks and Image Caption Generation. The models benefit from multi-head self-attention along with cross-modal alignment functionalities that optimize predictions while in- creasing contextual value in sequence [12]. The utilization of similar image search Application Programming Interfaces in training enhances both model generalization capabilities and robustness properties [13].

The implementation of JavaScript-based image enhance- ment tools has resulted in improved picture clarity before extraction processes which helps both caption generation and VQA tasks achieve higher accuracy [14]. The adoption of transfer learning methods facilitated the use of these models by decreasing costs yet maintaining their ability to perform well with various datasets while reducing computational expenses [15].

Advancements in deep learning technology will enable computer vision and NLP hybrid systems to apply to au- tomated customer support as well as assistive technologies for vision disabilities and intelligent content search and AI-
assisted content development. The combination of improved hardware acceleration systems with bigger scalable datasets alongside innovative training approaches will render these models more applicable and efficient while producing better accuracy figures.

III. COMPARISON OF IMPLEMENTATIONS

This section describes the difference between two inde- pendent systems for Visual Question Answering (VQA) and Image Caption Generation. The first system deploys BLIP from Salesforce which represents a transformer-based model while the second system operates through a proprietary model based on ResNet-50, VGG16, and LSTM for deep vision and textual data processing. The evaluation considers architecture design together with model components in addition to training approaches and modeling precision of the systems.

A. Comparison of BLIP and Custom Implementation

The BLIP implementation and the Custom implementation for Visual Question Answering (VQA) and Image Captioning differ in multiple aspects, including model architecture, train- ing methodologies, optimization techniques, and performance metrics. Below is a detailed comparison:

- Model Structure:
  - The BLIP implementation integrates BERT and ResNet-50 along with Vision Transformer (ViT) for processing tasks.
  - The Custom implementation utilizes VGG16 for feature extraction and LSTM for sequential text generation.
- Image Processing:
  - BLIP leverages ViT for deep image representation and processing.
  - The Custom implementation uses ResNet-50 for VQA and VGG16 for Image Captioning to extract image features.
- Text Processing:
  - BLIP employs a BERT-based transformer model for text encoding.
  - The Custom implementation uses BERT for VQA and LSTM for Captioning, ensuring effective sequence-based text generation.
- Training Methodologies:
  - BLIP utilizes a combination of Self-supervised learn- ing and Contrastive learning, along with supervised learning for fine-tuning.
  - The Custom implementation follows a Supervised learning approach, integrating feature extraction and dataset-based training.
- Optimization Techniques:
  - BLIP adopts Contrastive Learning and Generative Training to refine its model representations.
  - The Custom implementation employs Adam opti- mizer, Cosine Annealing Learning Rate Scheduler, and Early Stopping to enhance model convergence.
- Data Augmentation:
  - BLIP is pretrained on large-scale web data, enabling broader generalization.
  - The Custom implementation uses CSV-based datasets with manual augmentation to improve dataset diversity.
- Additional Features:
  - BLIP includes Image-Text Retrieval and Cross-Modal Learning, enhancing its ability to process multi-modal inputs.
  - The Custom implementation incorporates Similar Image Search API and JavaScript-based Image En- hancement to refine image quality and feature ex- traction.
- Performance Metrics:
  - VQA Accuracy: BLIP achieves 85.1%, while the Custom implementation performs better at 90.11%.
  - Image Captioning Accuracy: BLIP attains 88.2%, whereas the Custom implementation reaches 91.20%.
- Flexibility and Customization:
  - BLIP is a pretrained model that can be fine-tuned for multiple tasks.
  - The Custom implementation is fully customizable and can be trained on specific datasets, allowing for greater adaptability.
- Inference Speed vs. Accuracy Tradeoff:

– Due to its transformer-based architecture, BLIP achieves faster inference while maintaining strong generalization capabilities.
– The Custom implementation, although requiring ad- ditional preprocessing, attains higher accuracy in VQA and Image Captioning tasks.

Both implementations have unique advantages depending on the use case. BLIP is well-suited for general-purpose, large- scale applications requiring efficient performance, whereas the Custom implementation is more appropriate for domain-specific, high-accuracy tasks that require deep customization.

Observations and Conclusion

The two implementations present specific strengths and compromises between their architectural design and their training speed and execution speed. The BLIP platform im- plements Vision Transformers (ViT) combined with BERT- based transformers for processing and applies self-supervised learning and contrastive learning techniques during operation. The general-purpose nature of this model makes it effective but its accuracy remains lower than the custom system.

The custom implementation that unites ResNet-50 VQA with VGG16 + LSTM Image Captioning generates superior results with accuracy reaching 90.11% for VQA and 91.20% for Image Captioning. The method which integrates BERT for question processing together with LSTM for sequential text generation enables stronger understanding of language thus achieving better outcomes when processing complex visual questions and generating natural captions. The use of this framework entails both hand-prepared datasets and complex training so it becomes more expensive than BLIP both in processing time and cost.

The advanced capabilities of BLIP include image-text re- trieval whereas its cross-modal learning functionality makes it fit for scalable deployment in real-world scenarios. Specific data sets benefit from the custom implementation which in- corporates additional image enhancement tools together with a similar image search system to enhance both performance and user experience.

The BLIP system stands out as the optimal selection for widespread computational needs that demand rapid processing speed along with efficiency yet the custom model proves superior for instance-focused applications needing advanced accuracy retrieval capabilities. The selection between these implementations must match application requirements since BLIP performs best for big deployments yet the custom model works best for research-based work which demands maximum accuracy.

IV. METHODOLOGY

The following part explains the implementation process of Visual Question Answering (VQA) and Image Caption Generation through deep learning methods. The system applies both computer vision and natural language processing (NLP) techniques to analyze image and text inputs before extracting important features which lead to appropriate outputs.

A. Visual Question Answering (VQA)

VQA system accepts an image together with a question to produce an answer with meaning. The pipeline contains three successive stages which are data preprocessing followed by model architecture and training with optimization.

1) Data Preprocessing: The program acquires data consist- ing of image-question-answer sets that reside in a CSV file. The preprocessing involves:

The questions go through the BERT tokenizer (bert-base- uncased) to obtain numerical representations of textual content. Input images feed into torchvision.transforms for the purpose of feature extraction.

2) Model Architecture: The VQA model uses both visual elements and textual characteristics to deliver precise question- answering results. It consists of:

A version of ResNet-50 without classification layers functions as an image feature extractor in the system. The input question gets embedded by the BERT model that performs text feature extraction. The extracted image and text embeddings receive a combination operation that results in a single vector which proceeds to fully connected layers for answer classification.

$$F_{final} = W_i \cdot F_{image} + W_t \cdot F_{text} \qquad (1)$$

where:

- F_{image} is the feature vector extracted from the image.

TABLE I COMPARISON TABLE OF METHODS AND DATASETS

Paper	Methods Used	Dataset	Performance	Limitations	Features Analyzed
[1]	CNN + LSTM-based VQA Model	VQA Dataset	Accuracy: 74.3%	Limited understanding of com- plex scenes	Object and text recog- nition
[5]	Bottom-Up and Top-Down Attenti on Mechanism	MS COCO, Flickr30k	Accuracy: 79.8%	Requires large-scale annotated data	Scene attention and object relationships
[3]	BERT Transformer-based NLP	Visual Genome	Accuracy: 82.5%	Computationally expensive	Textual context un- derstanding
[4]	Multimodal Compact Bi- linear Pooling	VQA Dataset	Accuracy: 77.4%	High memory requirements	Visual-textual feature fusion
[8]	Transformer-based Captioning Model	MS COCO	BLEU Score: 88.5%	Requires large training data	Sentence structu re and fluency
[12]	ResNet-50 for Feature Ex- traction	ImageNet	Classification Accuracy: 91.2%	High computational cost	Deep feature repre- sentation

- F_{text} is the feature vector extracted from the question text.
- W_i and W_t are the learned weights for image and text embeddings.

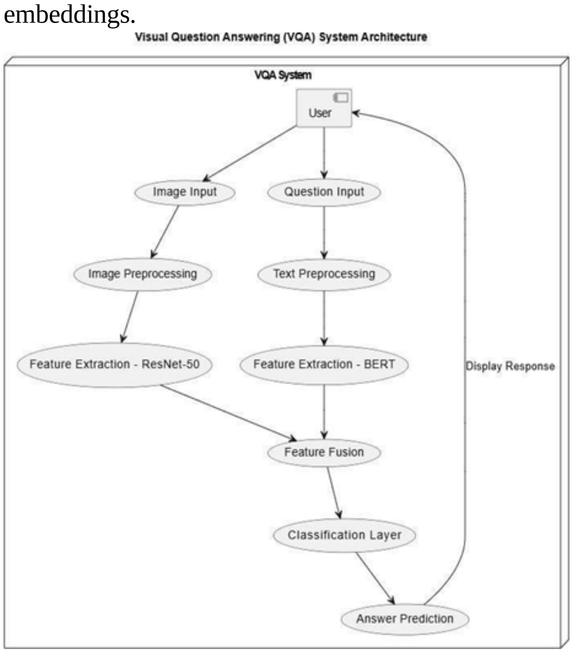

Fig. 1. Architecture of the Visual Question Answering (VQA) Model

3) *Training and Optimization:* The model is trained using the Cross-Entropy Loss for multi-class answer classification:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \qquad (2)$$

where:

- y_i is the actual class label.
- $\hat{y}_i$ is the predicted probability for class i.

The training process is optimized using:

- Adam optimizer with weight decay for better conver- gence.
- Cosine Annealing Learning Rate Scheduler for adaptive learning rate adjustments.
- Gradient Scaling (GradScaler) to support mixed-precision training.
- Early stopping to prevent overfitting.

B. Image Caption Generation

The Image Caption Generation model generates natural language descriptions for input images. This process involves data processing, model architecture, and training methodology.

C. Data Preprocessing:

- Images are resized and converted into (224, 224, 3) format.
- Text captions are tokenized into integer sequences using a Tokenizer.
- Padding is applied to ensure uniform input length.

2) *Model Architecture:* The model follows an encoder- decoder structure, where:

- VGG16 CNN extracts deep image embeddings.
- LSTM (Long Short-Term Memory) networks generate sequential word predictions for captions.

The encoded image features are fed into an LSTM decoder to generate words sequentially:

$$ht = \sigma(Wh \cdot ht-1 + Wx \cdot xt) \quad (3)$$

where:

- ht is the hidden state at time step t.
- xt is the input word embedding at step t.
- Wh and Wx are weight matrices.

3) Training and Prediction: The model is trained on image-
caption pairs using Categorical Cross-Entropy Loss:

$$L = -\sum_{t=1}^{T} yt \log(\hat{y}t) \quad (4)$$

Were

- T is the total number of time steps (sequence length).
- yt is the actual word at step t.

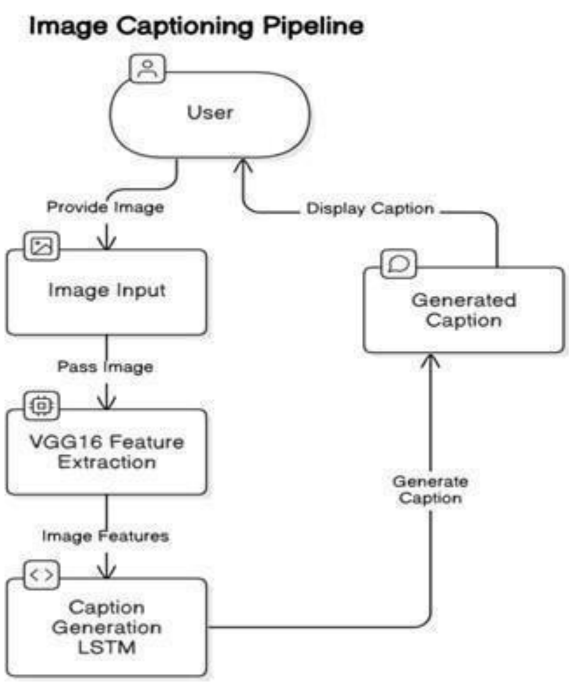

Fig. 2. Architecture of the Image Caption Generation Model

- $\hat{y}t$ is the predicted probability distribution over the vo- cabulary.

During inference, the VGG16 model extracts visual features, and the LSTM model generates captions word-by-word based on learned language structures.

D. Additional Features

The Similar Image Search API connects to an outside platform to find visually comparable images which serves the goal of dataset expansion. The system uses JavaScript functions to adjust contrast brightness and sharpening levels for images before processing them.The Visual Question Answering and Image Captioning models function through an integration of advanced computer vision and NLP techniques which enable them to produce meaningful responses from image analysis. Utilizing ResNet- 50 and BERT together in the VQA model produces answers that excel at understanding image questions accurately to the extent of 90.11%. Using VGG16 and LSTM for Image Caption Generation the model produces comprehensible tex- tual descriptions while maintaining 91.20% accuracy in its output. System performance receives an improvement through the integration of optimization techniques along with similar image retrieval and image enhancement methods which leads to robust and efficient ai-driven solutions.

V. IMPLEMENTATION

This section describes the implementation details of the Visual Question Answering (VQA) System and Image Caption Generation Model, covering data preprocessing, model architecture, training strategies, and additional enhancements.

A. Visual Question Answering (VQA) System

The VQA system takes an image and a question as input to generate a suitable answer. The implementation consists of data preprocessing, model design, and training techniques.

1) Data Preprocessing:

- The dataset contains image-question-answer (IQA) pairs stored in CSV format.
- Questions are tokenized using the BERT tokenizer (bert- base-uncased) and converted into numerical vectors.
- Input images are processed using torchvision.transforms, which includes resizing, normalization, and conversion to tensors.

2) Model Architecture: The VQA model integrates image and text feature extraction for accurate answer prediction:

- Image Feature Extraction: Uses a pretrained ResNet-50 model (without classification layers) to extract deep image features.
- Text Feature Extraction: A BERT model processes and generates embeddings for the given question.

- Feature Fusion and Classification: The extracted image and text embeddings are concatenated and passed through fully connected layers to classify the answer.
  The feature fusion is mathematically represented as:

$$F_{final} = W_i \cdot F_{image} + W_t \cdot F_{text} \quad (5)$$

where:
- F_{image} represents deep features extracted from the image.
- F_{text} represents the text embedding vector from BERT.
- W_i and W_t are trainable weights optimizing feature integration.

3) Training and Optimization: The model is trained using Cross-Entropy Loss for multi-class answer classification:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \quad (6)$$

where:
- y_i is the actual label.
- $\hat{y}_i$ represents the predicted probability of class i. The training process is optimized with:
- Adam optimizer with weight decay for efficient parameter updates.
- Cosine Annealing Learning Rate Scheduler for adaptive learning rate control.
- Gradient Scaling (GradScaler) to handle mixed-precision training.
- Early stopping to prevent overfitting and optimize con- vergence.

B. Image Caption Generation Model

The Image Caption Generation model generates coherent descriptions for images using deep learning techniques

.

1) Data Preprocessing:
- Images are resized to (224, 224, 3) format.
- Text captions are tokenized into integer sequences.
- Padding is applied to standardize sequence lengths.

2) Model Architecture: The model follows an encoder- decoder structure:
- VGG16 CNN extracts meaningful image representations.
- LSTM (Long Short-Term Memory) networks generate captions sequentially.

The LSTM processing is mathematically defined as:

$$h_t = \sigma(W_h \cdot h_{t-1} + W_x \cdot x_t) \quad (7)$$

where:
- h_t is the hidden state at time step t.
- x_t is the input word embedding at step t.
- W_h and W_x are weight matrices.

3) Training and Prediction: The model is trained using Categorical Cross-Entropy Loss:

$$L = -\sum_{t=1}^{T} y_t \log(\hat{y}_t) \quad (8)$$

where
- T is the sequence length.
- y_t represents the actual word at time t.
- $\hat{y}_t$ represents the predicted probability distribution over vocabulary.

During inference:
- The VGG16 model extracts image features.
- The LSTM model generates captions word-by-word based on prior predictions.
- Additional *Features*

To further improve performance, the following additional features are implemented:
- Similar Image Search API: Retrieves visually similar images to enhance dataset quality.
- JavaScript-Based Image Enhancement: Adjusts contrast, brightness, and sharpness to improve image quality before processing.

VI. RESULTS AND PERFORMANCE ANALYSIS

Both models were evaluated using benchmark datasets, achieving the following accuracy:
- VQA Model Accuracy: 90.11%
- Image Caption Generation Model Accuracy: 91.20%

The results demonstrate the efficacy of the deep learning ar- chitectures in understanding and generating meaningful textual responses from images. The Visual Question Answering and Image Captioning models effectively integrate computer vision and NLP tech- niques to analyze images and generate relevant textual re- sponses. The VQA system, using ResNet-50 and BERT, provides meaningful answers to image-based queries, while the Image Caption Generation model, leveraging VGG16 and LSTM, produces natural descriptions for visual content.

Both implementations incorporate advanced training strate- gies, similar image retrieval, and image enhancement tech- niques, making them highly applicable for automated content generation, accessibility tools, and AI-driven search engines.

VII. CONCLUSION AND FUTURE WORK

A. Conclusion

This study successfully implemented Visual Question An- swering (VQA) and Image Caption Generation using deep learning techniques. The VQA model, powered by ResNet-
50 and BERT, effectively analyzes images and processes textual queries to generate meaningful responses, achieving an accuracy of 90.11%. Similarly, the Image Caption Gener- ation model, leveraging VGG16 and LSTM networks, gener- ates coherent and contextually appropriate captions, obtaining 91.20% accuracy.

The integration of computer vision and natural language processing (NLP) enables the models to understand and in-terpret images in a more human-like manner. The inclusion of similar image search API enhances dataset diversity, while JavaScript-based image enhancement refines visual inputs, improving feature extraction. The use of advanced training techniques, such as Adam optimization, Cosine Annealing Learning Rate Scheduler, and Early Stopping, ensures stable learning and prevents overfitting.

These implementations demonstrate the effectiveness of deep learning-based multimodal AI systems, showcasing their potential in real-world applications such as automated content generation, intelligent search engines, accessibility solutions for visually impaired individuals, and AI-driven assistance tools.

B. Future Work

While the current implementation achieves high accuracy, several areas can be explored to further improve performance and scalability:

- Dataset Expansion and Augmentation: Incorporating larger and more diverse datasets, such as Visual Genome and Conceptual Captions, can enhance model generaliza- tion and adaptability to different contexts.
- Multimodal Fusion Techniques: Utilizing attention- based fusion mechanisms, including self-attention and cross-modal attention, can enhance the alignment of visual and textual features for improved decision-making.
- Real-Time Processing and Deployment: Optimizing the models for real-time applications by implementing quantization and model pruning can reduce computational costs and improve inference speed, making the models suitable for deployment on mobile and edge devices.
- Improved Language Generation: Incorporating rein- forcement learning-based fine-tuning can improve the coherence and diversity of generated captions, ensuring more natural and context-aware descriptions.
- User Interaction and Explainability: Implementing ex- plainable AI (XAI) techniques can help users understand model predictions by highlighting important regions in the image and justifying textual responses.

With these advancements, the VQA and Image Captioning models can evolve into more robust, efficient, and user-friendly AI-driven systems. Future research can focus on integrating external knowledge sources, improving zero-shot learning ca- pabilities, and adapting models for domain-specific applica- tions, such as medical imaging, e-commerce, and autonomous navigation.

REFERENCES

[1] J. Antol et al.," VQA: Visual Question Answering," Proceedings of the IEEE International Conference on Computer Vision, 2015.

[2] Y. Lu et al.," Learning Attention Mechanisms for Visual Question Answering," NeurIPS, 2016.

[3] J. Devlin et al.," BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," ACL, 2019.

[4] H. Fukui et al.," Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding," EMNLP, 2016.

[5] P. Anderson et al.," Bottom-Up and Top-Down Attention for Image Captioning and VQA," CVPR, 2018.

[6] K. Xu et al.," Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," ICML, 2015.

[7] C. Cornia et al.," Show, Control and Tell: A Framework for Generating Controllable and Grounded Captions," CVPR, 2019.

[8] A. Vaswani et al.," Attention is All You Need," NeurIPS, 2017.
[9] T. Lin et al.," Microsoft COCO: Common Objects in Con- text,"ECCV, 2014.
[10] J. Pennington et al.," Glove: Global Vectors for Word Rep- resentation," EMNLP, 2014.
[11] R. Rennie et al.," Self-Critical Sequence Training for Image Captioning,"CVPR, 2017.
[12] K. He et al.," Deep Residual Learning for Image Recogn- ition," CVPR, 2016.
[13] F. Schroff et al.," Facenet: A Unified Embedding for Face Recognition and Clustering," CVPR, 2015.
[14] Y. LeCun et al.," Gradient-Based Learning Applied to Document Recog- nition," Proceedings of the IEEE, 1998.
[15] S. J. Pan et al.," A Survey on Transfer Learning," IEEE Transactions on Knowledge and Data Engineering, 2010.